This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE Globecom 2010 proceedings.

On the Privacy of Encrypted Skype Communications

Benoît Dupasquier, Stefan Burschka, Kieran McLaughlin and Sakir Sezer
Queen's University of Belfast
Centre for Secure Information Technologies (CSIT)
Belfast, Northern Ireland, UK
E-mail: bdupasquier01@qub.ac.uk

***Abstract*—The privacy of voice over IP (VoIP) systems is achieved by compressing and encrypting the sampled data. This paper investigates in detail the leakage of information from Skype, a widely used VoIP application. In this research, it has been demonstrated by using the dynamic time warping (DTW) algorithm, that sentences can be identified with an accuracy of 60%. The results can be further improved by choosing specific training data. An approach involving the Kalman filter is proposed to extract the kernel of all training signals.**

I. Introduction

In the last few years, the evolution of voice over IP (VoIP) has been tremendous. The fact that it is free of charge, easy to use and apparently secure has promoted this technology among the population. Quickly, however, groups, with illegal or criminal intent, realized the potential of using VoIP as a covert channel. This has led several federal agencies, such as the National Security Agency (NSA) and the German Federal Police, and recently the European Union to demand the ability to eavesdrop on VoIP conversations. However, several service providers reject these demands without a legal warrant, to preserve privacy rights. Therefore, alternative methods, such as using domain specific knowledge to determine whether a given person uses sentences like *"I put the bomb in the train"*, have to be found to enable law enforcement to acquire an indication that might help to obtain a warrant for further investigations.

As VoIP becomes increasingly used for professional purposes, the need for most people think, an encrypted VoIP conversation does not completely prevent an attacker from eavesdropping. Using only side channel knowledge such as packet sizes or inter-arrival times, this paper demonstrates the false sense of privacy provided by Skype. To ensure the validity of the proposed results, more than 6 GB of Skype traces have been collected as training and test material. This paper investigates how an eavesdropper can determine the content of an encrypted VoIP conversation. This goal is achieved by identifying particular sentences in the flow of Internet Protocol (IP) packets using algorithms from speech processing, such as dynamic time warping (DTW), and from radar tracking analysis, such as the Kalman filter.

This paper is organized as follows: some related work concerning VoIP security and encryption is briefly presented in Section I-A. The motivation is explained in Section I-B. The methodology is presented in Section II, including the experimental setup and the requisite background for understanding the algorithms proposed. The results are discussed in Section III. Finally, the conclusion is presented in Section IV.

A. Related Work

Literature shows that privacy is not yet fully guaranteed with VoIP, as illustrated by number of published papers, ranging from inferring speech activity to identifying sentences or the language of the conversation.

A traditional speaker identification system usually has to decompress the compressed stream before being able to analyze it. Aggarwal et al. [1] show that decompression can be avoided. This challenging goal is achieved by using a so-called *micro-clustering algorithm*, detailed in [2]. They reach an accuracy of 80% with a three times higher speed than the Gaussian mixture model (GMM), used in the traditional approach. Moreover, they are able to recognize the speaker in real time after five seconds of seeing the first packet in a stream.

In order to maintain user datagram protocol (UDP) bindings at the network address translation (NAT) and to obtain better voice quality, Skype does not suspend sending packets during silence. This means that silence is encoded, encrypted and sent over the network, therefore complicating the task of inferring speech activity, a process commonly called voice activity detection (VAD). Nevertheless, Chang et al. [3] propose a method they named network-level VAD to infer speech activity from encrypted voice traffic. Their work is based on the fact that speech activity is highly correlated with the size of encrypted packets, i.e., more information is encoded in a voice packet while the user is speaking than while the user is silent. In their paper, they devise an adaptive thresholding algorithm allowing silence to be differentiated from voice activity with an accuracy of 85%.

In their paper *Privacy of Encrypted Voice-over-IP* that focuses on Google Talk, Lella and Bettati [4] assume that the packet sizes do not carry any information but that all the information is contained in the inter-arrival times. They use the silence phases, which are encoded in smaller packets and with bigger inter-arrival times, to isolate words. Then, they compare the length of the words, namely the timestamp of the last packet belonging to the word minus the timestamp of the

first one, against their models. Their first approach, based on a simple Bayesian classifier, is context unaware. Subsequently, they propose a second approach based on hidden Markov models (HMMs) to produce context awareness. A drawback to their approach is that a congested network or a slow or lossy link will render their attack ineffective. Moreover, it relies on human analysis to determine the plausibility of the proposed sentence.

Wright et al. [5] propose to use the size of transmitted packets to determine whether a given phrase is spoken. By reducing speech to phonemes, the most basic speech unit, they are able, using HMMs, to reconstruct sentences using the sequence of packet lengths. By reducing the problem of encrypted VoIP speech recognition to a substring matching problem on the packet sizes, the authors are able to achieve an average accuracy of 50% in telling whether a sentence is contained in a stream of packets. By choosing some phonetically rich sentences, such as *"Young children should avoid exposure to contagious diseases"*, they achieve an accuracy above 90%. By using a similar approach and a variant of the χ^2 test, the authors have also demonstrated that determining the language of an encrypted conversation is possible [6]. Their algorithm achieves an accuracy of 66% in identifying the language of a conversation among a set of 21 models. Furthermore, 14 languages can be identified with an accuracy greater than 90%. Finally, their algorithm achieves an overall binary classification rate of 86.6% However, a significant limitation to their approach is that it needs the binary of the audio codec used and requires the VoIP application to encode voice using a variable bit rate (VBR) codec and a length preserving encryption mechanism.

To conclude, many experiments have been conducted on this topic that do not take account of the time dependency of the packet signals. The work presented in this paper will now show that it is a factor with information gain even without the availability of the codec.

B. Motivation

This paper addresses the challenge of extracting information from encrypted VoIP traffic. Although several studies have already tackled this issue, none of them have tried to attack Skype or other closed source codecs. This represents a real challenge as the measurement and process noise have to be taken into account. This requires the investigation into preprocessing methods able to remove, attenuate or take the effect of noise into account.

The efficiency of the approach presented in [4] is strongly dependent on the quality of the network. Furthermore, sentences of similar lengths cannot be reliably identified. This paper overcomes this restriction by focusing on packet sizes rather than inter-arrival times. In addition, it is proposed that by using the DTW algorithm presented in Section II-B, variability in time or speed can be efficiently dealt with. Furthermore, unlike HMMs, the DTW algorithm does not require an extensive amount of training data to work properly. This means that a single occurrence of a sentence spoken by a known speaker might theoretically be sufficient to generate a model for this particular speaker.

Experiments have shown that the approach proposed in [5] does not work with Skype traffic as the encoding of phoneme depends strongly on the surrounding sounds. Furthermore, the approaches proposed in [5] and [6] require the possession of a binary of the audio codec to create a match between the VBR output and the encrypted packet length. However, several audio codecs are closed source and do not share their binary. Without a binary, such a mapping could not be created. The non-availability of the codec renders the attack more difficult, and this clearly creates an issue that cannot be solved by the approach proposed in [5] or [6]. Indeed, to investigate unknown or closed source codecs, methods that eliminate the influence of the encryption algorithm have to be developed. The strength of the methodology presented in this paper compared to the referenced techniques is that it can be applied to any VoIP application, regardless of the availability of the codec's binary.

II. Methodology

All experiments are conducted on Skype version 4.0.0.224 working in direct UDP mode. The laboratory setup is composed of two Microsoft Windows XP workstations. This choice is lead by the desire to work on the latest technologies offered by Skype, mainly its new audio codec SILK and to provide results valid for as many users as possible. Some prerecorded speech is played on one side, while all the traffic between the two stations is captured using *WinDump* [7].

To facilitate the analysis process, Skype is operated in direct UDP mode and in a noise-free environment. Furthermore, synthetic speech, being close to the nature of Skype's vocoder, is being used. Finally, several recordings of the model are created. By observing IP traces of Skype conversations, it has been determined that information is carried in the time-dependent flow of packet lengths. The number of packets and the amount of bytes transferred have been identified as features relating to a given sentence. Perceiving the sequence of packets as a signal inspired the application of methods from speech processing or radar signal analysis.

An extensive analysis of the packet sizes used by encrypted Skype traffic revealed that a given sentence is always encoded in a similar way, and conversely, different sentences produce different outputs as illustrated in Figure 1. This is an important observation as it implies that, in order to determine the most probable model, an algorithm able to compare sequences of packets is sufficient. The DTW algorithm has been chosen for this purpose.

In order to avoid the generation of several speaker models per sentence, the Kalman filter is used toward the generation of speaker-independent models. To begin with, the first observation is chosen as the initial estimator of the state of the system, and then, the Kalman filter is recursively applied to the training data. Finally, the DTW algorithm is used to compare the test data against the Kalman's models.

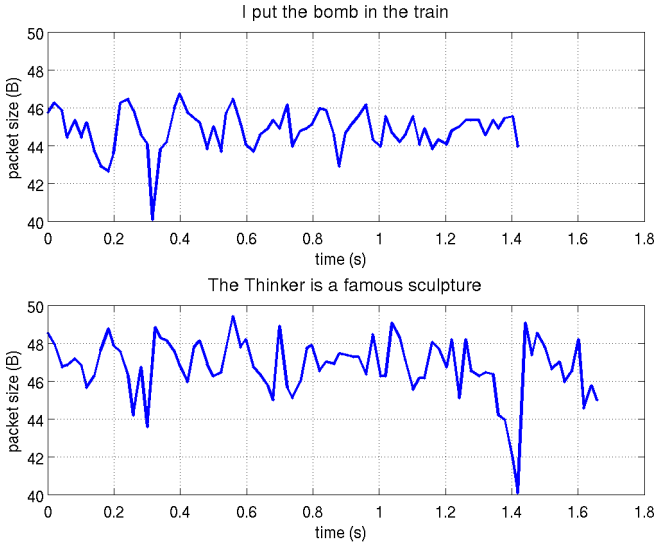

Fig. 1. Different sentences, different outputs

A. Limitations and Assumptions

All the experiments are conducted using the English language, since it is the most common one on the Internet. Nevertheless, this work can be adapted to analyze streams of any other language.

Although it is focused on Skype, this work can be adapted to any other existing VoIP software. Skype works independently of the network topology and security measures. In order to focus on the vital part, the content appraisal from IP packets, this work concentrates first on the ideal case without a firewall or NAT. This also prevents too many unknown variables from influencing the experiments. To simplify the experiments and in order to understand the fundamental principles behind Skype traffic, this work focuses on an idealistic case. First, it is assumed that the direct UDP mode is used. This setup has been chosen because it is the one that induces the least interferences, e.g., delays and losses. However, it is not the most common in the Internet.

Second, samples of prerecorded synthetic speech are used, which do not vary between the different experiments. In practice, speech is dynamic, and a sentence uttered by the same speaker will not always sound exactly the same way. Third, a conversation is only defined by the audio traces of Skype, i.e., no text messages, file transfer or video embedded.

B. Background

In this section, the algorithms relevant to the present attack method are briefly presented. The two key algorithms used are dynamic time warping (DTW) and the Kalman filter. Several references are provided for anyone requiring a deeper understanding of those algorithms.

1) Dynamic Time Warping (DTW): DTW is a dynamic programming algorithm allowing sequences, possibly of different length, to be compared. Therefore, DTW is an algorithm suitable for comparing two sequences that may vary in time or speed. It has been extensively used, especially in speech processing before being substituted by HMMs.

The result of the DTW algorithm comes in many forms. The first one is a value representing the distance, i.e., the correlation factor, between both sequences. The closer to zero this distance is, the more similar the input sequences are. It also supplies a graphical representation, displaying a matrix of size $m \times n$, where the abscissa represents one sequence and the ordinate axis the other one. Each cell (i, j) of the matrix is colored according to the distance $dist(x_i, y_j)$, e.g., the Euclidean distance between x_i and y_j. The "colder" the color, the smaller the distance. The optimal DTW path is marked by a white line. A perfect match describes a straight line from the upper left of the matrix to the bottom right, namely the point (m, n). More details on the DTW algorithm can be found in [8] and [9].

2) Kalman Filter: The Kalman filter is a recursive linear filter that estimates the state of a linear, discrete-time dynamical system from a series of noisy measurements. Basically, the goal of the algorithm is, given the current state of the system, to predict the next state and its uncertainty. Then, the prediction is corrected with the new measured values. It is mainly used for radar or missile tracking problems but is also widely used in computer vision, economics and navigation (aerospace, land and marine). The filter combines all available measurement data, plus prior knowledge about the system and measuring devices and can be used to adaptively filter and predict a discrete-time linear process. Moreover, it can be shown that the Kalman filter is the optimal linear filter, namely the one that minimizes most the variance of the estimation error. Details on the original proposals for using Kalman filters can be found in [10] and [11], while more recent work and development can be found in [12].

III. Results

As mentioned in the previous section, a first encouraging observation is the coherence and reproducibility of the experiments. By analysing the sequence of IP packets generated for an experiment run several times under the same conditions, only minor changes can be observed. This proves that Skype does not generate random packet sizes as the output appears to be deterministic.

In order to evaluate sentences, reference models are required for comparison with the recordings under examination. Experiments have shown that the inter-arrival time evolves from 60 ms to 40 ms to finally reach 20 ms. This behavior appears to be due to the adaptive nature of the codec learning to differentiate silence and noise from speech. Therefore, the models are generated by a WAVE file consisting of three occurrences of a sentence separated by 30 seconds of silence, so sentences can easily be extracted from the training data.

Then, labeled data is generated, namely several speech samples are recorded in which the given sentence appears in different positions. By manually extracting the given sentence from the traces, the DTW algorithm, presented in Section II-B1, can successfully be used to detect the sentence. Figures

2 and 3 show the graphic results of the DTW experiment for a non-matching sentence and a matching sentence, respectively. Recall that, as explained in Section II-B, a perfect match would result in a straight line through the diagonal. The smaller graphs on the left and the bottom of each DTW graph represent the input sequences, respectively, the model and the test data.

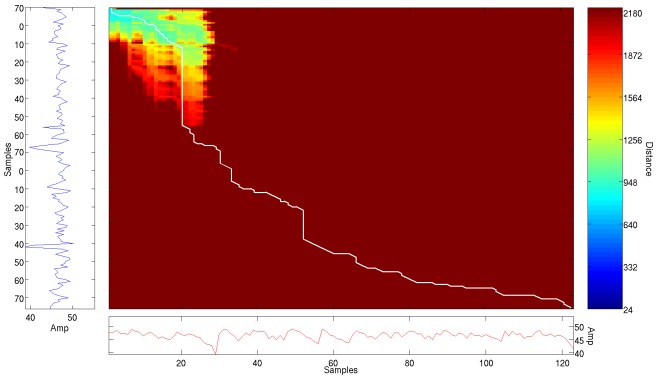

Fig. 2. A non matching DTW path

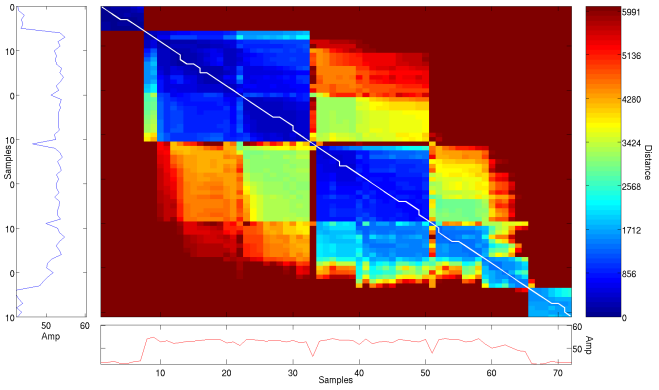

Fig. 3. A matching DTW path

In order to evaluate the efficiency of the content appraisal attack proposed, several models are generated for sentences chosen to illustrate the ability to detect and distinguish even similar sentences from encrypted Skype traces. Variations of the sentences *"I put the bomb in the plane"* were used, as they present an obvious interest to law enforcement. The other sentences have been chosen for their length or the similarity of the content to this initial sentence. Finally, the following test set is used as an input to the sentence detector:

1) Although always alone we survive,
2) Biblical scholars argue history,
3) I put the bomb in the bus,
4) I put the bomb in the plane,
5) I put the bomb in the tower,
6) I put the bomb in the train,
7) The bomb is in the train,
8) The Thinker is a famous sculpture,
9) There is a bomb in the train.

The test set is composed of 6 experiments, namely of all the possible permutations of sentences 1, 2 and 6. This allows the testing of the models against 18 samples (namely 3 sentences * 3! permutations). By creating models and extracting sentences manually, two sentences out of three (66%) have been correctly identified. If model 3 is removed, then an accuracy above 70% can be reached. It is also interesting to note that the results could have been improved by using different models and test data. For example, in the case $(2, 1, 6)$, 6 was identified as 3. If model 3 is removed, the sentence is correctly identified. This can be explained by the similarity of the sentences 3 and 6, differing only in the last word. Furthermore, the DTW algorithm has also been tested against sentences longer than the model, and as expected, each time the resulting distance was too big to be accepted as a match. This means that the recall and precision of the method can easily be improved by generating models for sentences that differ in duration.

A given sentence is not always encoded exactly the same way; therefore, by generating several models per sentence, an 83% correct identification rate has been achieved. However, the need to generate several models per sentence is not practical as the number of different models is potentially unbounded due to the variability of speech and noise. Therefore, in the next section, the use of the Kalman filter is proposed in order to extract the kernel of all signals.

A. Speaker Independent Models Using the Kalman Filter

The effect of the Kalman filter, introduced in Section II-B2, on four occurrences of the sentence *"Although always alone, we survive"* is illustrated in Figure 4. Although not obvious in the figure, a problem faced with this approach is that the training and test data do not always have the same length. This is particularly true between different speakers. Indeed, the model will be as long as the longest sample in the training set. Depending on the difference in size between the sample and the model to be compared against, there is a chance of misclassifying the sentence. Therefore, a slightly different approach is taken. The idea is not to compare the model against the sample, but rather to measure the distortion of the model when updated with the sample.

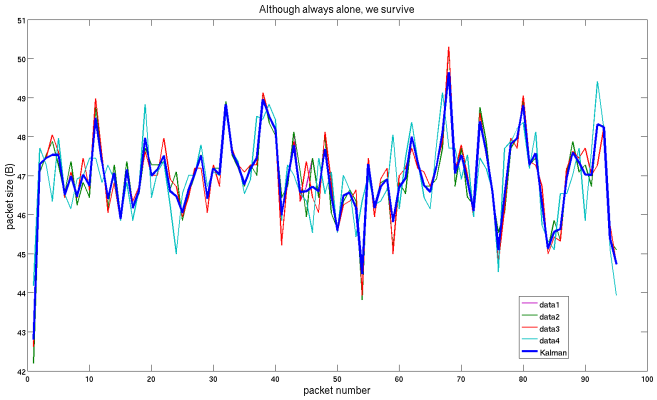

Fig. 4. Model generation using the Kalman filter

Let m be the model generated by the Kalman filter and s the test data to compare against the model. Then, the Kalman model is updated with the test data: $k' = Kalman(m, s)$. This means that m is the model and k' is the model corrected with the test data. Now, the distance between m and k' is computed with DTW, as it was done before without Kalman.

Matching sentences should not distort the model too much, while non-matching sentences should induce more variations. However, the problem observed with this approach is that the Kalman filter does not weight the sample enough. Therefore, no matter how different the test data and the model are, the distortion was minimal and the DTW algorithm inefficient. To approach this problem, the input parameters, such as the covariance of the estimation error, were adjusted. The values were selected in the interval of $(0, 1]$, but none of them produced satisfying results. This can be explained by the assumptions inherent to the Kalman filter, which assumes that both the process and the measurement noise are additive, white and Gaussian with zero mean. In the case of Skype, the measurement and process noise are unknown. Therefore, the assumptions made by the Kalman filter might lack the precision necessary to describe the actual noise faced by Skype.

The primary objective of the presented research was the demonstration (proof of concept) that significant information is leaked from Skype conversations. Therefore, further optimization of the above approach is considered as future research on speaker-independent speech recognition. Albeit there is a limited amount of testing and a small set of targeted sentences, this research work highlights the fact that extraction of information from encrypted VoIP conversations can be achieved without relying on the availability of public information on the codec, encryption, etc. to attempt a successful side channel attack.

IV. Conclusions

This paper has demonstrated the false sense of privacy provided by Skype, a widely used VoIP application, known for its strong security policy. Through a comprehensive analysis of encrypted Skype traffic, it has been shown that isolated phonemes can be classified and given sentences identified with an accuracy greater than 60%. An accuracy of 83% can be reached under specific conditions. This challenging goal is achieved with the help of the DTW algorithm, commonly used in the speech processing community. The algorithm is successfully applied on vectors of packet sizes, allowing the probability of a recording being a particular sentence to be determined. Furthermore, the Kalman filter, traditionally used for radar tracking problems, has been proposed to extract the kernel of the training data in order to generate speaker independent models.

To the knowledge of the authors, this is the first published research that demonstrates that Skype encryption is not entirely secure and that information is leaked allowing content to be inferred. In addition, dealing with one of the most challenging VoIP software constructs currently available and proposing a methodology easily adaptable to any other VoIP software means the proposed approach is very versatile. Furthermore, by using the DTW algorithm, the drawbacks of an approach involving HMMs, e.g., an extensive amount of training data, can be avoided. Also, the capacity of DTW to deal with time or speed variability along with the use of packet sizes as discriminant features renders the proposed approach stronger than the one described in [4]. Finally, because Skype is closed source and its audio codec is not widely published, the approach proposed in [5] and [6] could not have been applied. Therefore, it has been necessary to study alternate methods able to attenuate or take the effect of noise into account.

Another interesting study to conduct would be to design a speaker identification or verification system. Speaker identification addresses the problem of determining who is talking from a set of known speakers, while speaker verification is concerned with the problem of verifying that the speaker is really the person they purport to be. Any other profiling information might also be of interest, especially for law enforcement agencies. So far, only the language identification issue has been addressed [6], although not for Skype. However, based on the results of this paper, age, gender, language or emotion detection also seems feasible for Skype.

References

[1] C. C. Aggarwal, D. Olshefski, D. Saha, Z.-Y. Shae, and P. Yu, "CSR: speaker recognition from compressed VoIP packet stream," in *Proceedings of the IEEE International Conference on Multimedia & Expo*. Los Alamitos, CA, USA: IEEE Computer Society, Jul. 2005, pp. 970–973.

[2] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in *Proceedings of the 29th International Conference on Very Large Data Bases*. Morgan Kaufmann, Sep. 2003, pp. 81–92.

[3] Y.-C. Chang, K.-T. Chen, C.-C. Wu, and C.-L. Lei, "Inferring speech activity from encrypted Skype traffic," in *Proceedings of IEEE Globecom*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2008.

[4] T. Lella and R. Bettati, "Privacy of encrypted voice-over-IP," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. IEEE Computer Society, Oct. 2007, pp. 3063–3068.

[5] C. V. Wright, L. Ballard, S. E. Coull, F. Monrose, and G. M. Masson, "Spot me if you can: uncovering spoken phrases in encrypted VoIP conversations," in *Proceedings of the 29th IEEE Symposium on Security and Privacy*. IEEE Computer Society, May 2008, pp. 35–49.

[6] C. V. Wright, L. Ballard, F. Monrose, and G. M. Masson, "Language identification of encrypted VoIP traffic: Alejandra y Roberto or Alice and Bob?" in *Proceedings of the 16th USENIX Security Symposium*. Berkeley, CA, USA: USENIX Association, 2007, pp. 1–12.

[7] WinDump, "Official website," 2010, http://www.winpcap.org/windump.

[8] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in *1st SIAM International Conference on Data Mining*. SIAM, Apr. 2001.

[9] S. Salvador and P. Chan, "FastDTW: toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[10] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transaction of the ASME, Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, Mar. 1960.

[11] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Transactions of the ASME, Journal of Basic Engineering*, vol. 83, no. Series D, pp. 95–107, 1961.

[12] M. S. Grewal and A. P. Andrews, *Kalman Filtering: Theory and Practice Using MATLAB*, 3rd ed. John Wiley & Sons, 2008.